ABSTRACT

Language or parsing model is the heart of any NLP application like speech recognition and machine translation which provides ways and means for language analysis. Language model is the probability distribution over a sequence of strings in the form of n-grams. Language model is the mechanism which assists in generating output sentences from the sequence of words generated by any NLP application. Parsing is the process of analyzing text automatically by assigning syntactic structure according to the grammar of a natural language. A parser is a computational system which processes input sentences and builds one or more constituent structures called parses or parse trees. Parsing models are created by adding structural components in statistical approach which provide probability distribution over parses.

Tamil is a resource deficient language in terms of text and annotated corpora, tools and techniques. With the available corpus, an initial domain independent language model is developed. In order to overcome the deficiency of text corpus and probability sparseness in the language model, translation lexicons and their probabilities are generated in Tamil from English by means of Cross-lingual Latent Semantic Analysis (CLSA) using English-Tamil document aligned corpora. Partial morphology is applied to bridge the gap between Tamil and English words due to the inflectional nature of Tamil. Translation probabilities are adapted directly to the language model through interpolation technique. In order to improve probabilities of content
words related to various topics, topic specific models are developed and improved with translation lexicons obtained from CLSA after the identification of topics. Then, a domain independent language model is adapted with topic specific models. It satisfies factors like coverage, accuracy, simplicity and robustness to a greater extent.

Parsing models are trained with a large volume of annotated corpus called Treebank for satisfying factors like long distance relationship among words, free word ordering and coverage of syntax with semantics. Phrase structure parsing model covers the factors to a certain extent whereas dependency structure parsing model covers the same to a greater extent. To develop a phrase structure Treebank, morphological analysis, POS tagging and phrasing are needed. To develop a dependency structure Treebank, morphological analysis, POS tagging, and marking and labelling of relations are needed.

Tamil has a well-defined set of morpho-syntactic rules. A large extent of agglutination in Tamil leads to longer words with many combination of affixes. A Tamil word can have one prefix and seven suffixes at the maximum. But there are many exceptions and complex patterns. Tamil is a relatively free word order language. Ambiguity is also a problem in differentiating a noun and a verb, a verbal noun and a verb, and similar suffix patterns for different parts of speech. In order to make a POS tagger to learn morpho-syntactic rules with Transformation Based Learning and Hidden Markov Model techniques with the aforesaid problems, a large volume of
annotated corpus is needed. Hence the rule based approach is preferred for morphological analysis and POS tagging in Tamil.

There are several ways of developing the Treebank in Tamil. Firstly, a small initial hand annotated corpus is prepared and bootstrapping is applied with many cycles using manual corrections. Secondly, a rule based morphological analyzer, POS tagger and phrasing tool are created and employed. Thirdly, the rule based tools are improved with statistical methods like POS projection and morphological induction through sentence aligned corpora and employed.

The rule based morphological analyzer, POS tagger and phrasing tool have been successfully created in Tamil. Even though the rule based morphological analysis in Tamil is done with a reasonable accuracy using suffix patterns and partial stems, further improvement is difficult due to the non-availability of a standard method for finding root words and disambiguating the noun and the verb. The rule based morphological analyzer and POS tagger in Tamil have been improved with categorical information and root word, generated from English using POS projection and morphological induction respectively. Bootstrapping with hand annotation is time-consuming, laborious, and inconsistent. It also needs human expertise. Hence an improved rule based approach is employed for the generation of Treebank.

In this research, domain independent and topic specific language models have been developed in Tamil. The domain independent language
model is improved with direct and topic specific models adaptations with perplexity values of 118.04 and 115.56 respectively. Thorough morphological analysis has been done for the development of parsing model in Tamil. POS, phrasal and dependency tag-sets and phrasing rules have been proposed for Tamil. The rule based morphological analyser and POS tagger, and phrasing tool are developed with 85.56% and 75.11% accuracies respectively. The rule based morphological analyzer and POS tagger are improved to 92.48% with hybridization of statistical methods. Phrase and dependency structure Treebanks are developed with 10,000 sentences. Phrase and dependency structure parsing models are built using immediate head parsing and non-projective dependency parsing techniques with 95% and 96.65% accuracies respectively.